4A

УДК 61:004.651(075.8)

І.Є. Андрущак, Ю.С. Повстяна
Луцький національний технічний університет, Україна
Україна, 43000, м. Луцьк, вул.Львівська, 75

Програмна реалізація методу індукції дерева рішень для класифікації політравм: питання обчислюваної складності

I.Ye. Andruschak, Y.S. Povstiana
Lutsk National Technical University, Ukraine
Ukraine, 43000, Lutsk, Lvivska st. 75

Software Implementation By Induction of Decision Trees for Classification Of Polytrauma: Computational Complexity

И.Е. Андрущак, Ю.С. Повстяная
Луцкий национальный технический университет, Украина
Украина, 43000, г. Луцьк, ул. Львовская, 75

Программная реализация метода индукции дерева решений для классификации политравм: вопросы вычислительной сложности

В работе разработан и программно-реализован метод индукции дерева решений для задачи классификации политравм на основе ряда биохимических показателей. Изучаются вопросы вычислительной сложности алгоритма. Проект реализован в среде Netbeans на основе Java-классов.
Ключевые слова: политравма, принятия решений, дерево решений, Java, SQL.

In this work was developed and implemented a method of inducing decision trees for classification problem polytrauma based on a number of biochemical parameters. The problems of computational complexity of the algorithm. The project was implemented in an environment based on Netbeans Java-classes.
Keywords: polytrauma, decision-making, decision tree, Java, SQL.

У роботі розроблено та програмно реалізовано метод індукції дерева рішень для задачі класифікації політравм на основі ряду біохімічних показників. Вивчаються питання обчислювальної складності алгоритму. Проект реалізований в середовищі Netbeans на основі Java-класів.
Ключові слова: політравма, прийняття рішень, дерево рішень, Java, SQL.

Вступ

Під політравмою мають на увазі складний патологічний процес, зумовлений пошкодженням кількох анатомічних областей або сегментів кінцівок. Проблему становить правильне та своєчасне діагностування політравм, особливо в умовах надзвичайних ситуацій або військових дій, коли пацієнт знаходиться у стані без свідомості.

Метою даної роботи є розробити і програмно реалізувати алгоритм класифікації політравм з використанням методу індукції дерева рішень, вивчити питання його обчислювальної складності. Вирішувана проблема належить до широкого класу задач диференціальної діагностики. В медицині поняття «диференціальної діагностики» означає системний підхід, що грунтується на доказовості, для визначення причини симптомів, що спостерігаються, у випадку, коли є кілька альтернативних пояснень, а також для зменшення переліку можливих діагнозів.

Одним з підходів, що відображає природній процес мислення при диференціальній діагностиці, є метод індукції дерева рішень. У попередніх наших дослідженнях даний алгоритм було описано та програмно реалізовано в середовищі Netbeans для бази даних MySQL. Інтерес викликає проблема обчислювальної складності алгоритму для реальних клінічних даних – таких, як наприклад, дані біохімічних досліджень у випадку політравм.

Програмна реалізація. Метод реалізовано в середовищі розробки Netbeans на мові програмування Java. Базу навчальних даних розгорнуто на сервері MySQL. На рис. 1 представлено концептуальну модель інформаційної системи. У класі DecisionTree безпосередньо реалізовано метод індукції дерева рішень. У клас DataManager надходять виклики від DecisionTree на виконання запитів до бази даних mysql щодо отримання навчальних даних.

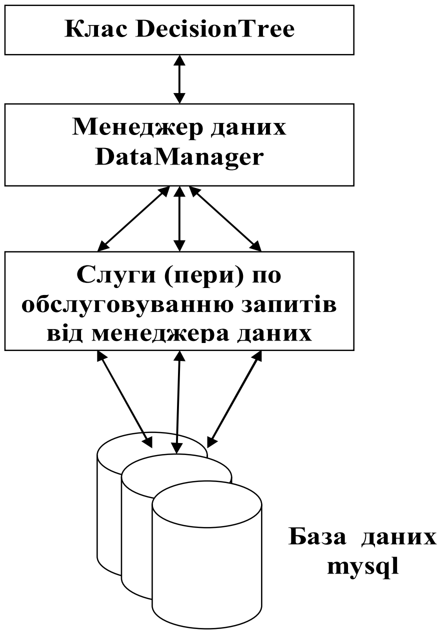


Рисунок 1 – Концептуальна модель інформаційної системи індукції дерева рішень

База даних mysql складається з двох таблиць – таблиці attribute, призначеної для зберігання інформації про атрибути та таблиці categorized_data – для наборів навчальних даних. Структура таблиць на мові SQL для класифікації політравм наведена нижче:

```
CREATE TABLE mysql.attribute (
        id integer not null unique,
    attribute_name varchar(25),
    attribute_field_name varchar(25),
  primary key (id)
 ) ENGINE=InnoDB;
CREATE TABLE mysql.categorised_data (
        id integer not null unique,
    A1 varchar(12),
    A2 varchar(8),
    A3 varchar(7),
      …………………..
    A21 varchar(7),
    class varchar(28),
  primary key (id)
 ) ENGINE=InnoDB;
```

Програмні класи проекту включено до пакету decision_tree.model. Сюди входять beans-класи Attribute, Attribute_for_list та CategorisedData для роботи з даними відповідних таблиць. SQL-запити щодо отримання відповідних даних, включаючи розрахунки інформаційних показників реалізовано в класі AttributeListPeer.

Клас DecisionTree є нащадком класу DefaultTreeModel пакету javax.swing.tree. Він має два елементи класу: m_dataManager – менеджер даних та m_htAttribute_list – хеш-таблиця із списком атрибутів. Хеш-таблиця із списком атрибутів (у методах класу DecisionTree виступає під назвою htAttribute_list) створюється для кожного вузла дерева рішень. Вона має два призначення – поряд із списком включених для даного вузла атрибутів зберігати умови поділу (splitting conditions), які перейшли до даного вузла від вузлів-батьків. Кожен вузол дерева рішень є об’єктом класу DefaultMutableTreeNode. В якості об’єкта кожен вузол зберігає об’єкт класу NodeObject, декларація якого наведена нижче:

```
class NodeObject {
   Attribute attribute;
   Hashtable htAttribute_list;
   String splitting_criterion;
   String sLabel;
   public String toString() {
      if (splitting_criterion.matches("")) { return sLabel; }
      else return "if '" + splitting_criterion + "' then '" + sLabel + "'"; }
}
```

Тут attribute – атрибут, який повертається методом Attribute_selection_method, splitting_criterion – умова поділу, яка переходить від батьківського вузла, sLabel – надпис на вузлі. Хеш-таблиця htAttribute_list використовується для побудови наборів навчальних даних D_j для кожного із вузлів і має таку структуру:

Тип ключа	int
Тип об’єкта	Attribute_for_list
Структура об’єкта	Attribute attribute; Hashtable htSplitting_outcomes; String splitting_criterion; boolean included;

Тут included – булева змінна-прапорець належності атрибуту attribute до списку атрибутів даного вузла. Можна показати, що коли included=true, то вузол з назвою attribute є для даного вузла дочірнім (на певному нижчому рівні ієрархії). У випадку, коли атрибут attribute не входить до списку атрибутів для даного вузла (included=false), то вузол з назвою attribute є батьківським (на певному рівні ієрархії), а в змінній splitting_criterion зберігається умова поділу, якій підлягає даний вузол відносно батьківського вузла attribute.

Хеш-таблиця htSplitting_outcomes містить усі можливі наслідки (умови поділу) щодо атрибуту attribute.

Метод Generate_decision_tree є безпосередньою реалізацією методу індукції дерева рішень. Заголовок методу має вигляд:

private DefaultMutableTreeNode Generate_decision_tree (Hashtable htAttribute_list, DefaultMutableTreeNode dmtnSubroot, String splitting_criterion)

В якості аргументів метод використовує кореневий вузол дерева, список пов’язаних з ним атрибутів htAttribute_list та умову поділу splitting_criterion. В якості значення метод повертає дочірній вузол типу DefaultMutableTreeNode. Шляхом рекурсивного виклику методу Generate_decision_tree будується дерево рішень.

З метою візуалізації представлення дерева використано клас javax.swing.JTree. При цьому дерево рішень створюється за допомогою операторів:

dtDecision_tree = new DecisionTree(dmtnRoot, dataManager, htAttribute_list);

jTree1.setModel(dtDecision_tree);

SQL-реалізація розрахунку інформаційних показників. Далі за допомогою розроблених програмних класів побудуємо дерево рішень щодо класифікації політравм на основі даних 6-ти класифікаційних груп, а саме:

- черепно-мозкова та скелетна травми мали місце 2 години тому;
- черепно-мозкова та скелетна травми з кровотечею мали місце 2 години тому;
- черепно-мозкова та скелетна травми мали місце 12 годин тому;
- черепно-мозкова та скелетна травми з кровотечею мали місце 12 годин тому;
- черепно-мозкова та скелетна травми мали місце 24 години тому;
- черепно-мозкова та скелетна травми з кровотечею мали місце 24 години тому.

Використано таку таблицю атрибутів:

INSERT INTO mysql.attribute (id, attribute_name, attribute_field_name) VALUES (1, 'Mass', 'A1'), (2, 'Total.bil.', 'A2'), (3, 'AsAT', 'A3'), (4, 'AlAT', 'A4'), (5, 'GP', 'A5'), (6, 'GR', 'A6'), (7, 'VG', 'A7'), (8, 'SOD', 'A8'), (9, 'Catalaza', 'A9'), (10, 'MDA', 'A10'), (11, 'DK', 'A11'), (12, 'TsP', 'A12'), (13, 'TsIK', 'A13'), (14, 'API', 'A14'), (15, 'Ig A', 'A15'), (16, 'Ig M', 'A16'), (17, 'Ig G', 'A17'), (18, 'Il-2', 'A18'), (19, 'Il-6', 'A19'), (20, 'Il-10', 'A20'), (21, 'TNF-a', 'A21');

Набори включають лише категоріальні дані (попередньо оброблені), наприклад:

INSERT INTO mysql.categorised_data (id, A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19, A20, A21, class) VALUES
(1, 'low', 'low', 'high', 'high', 'normal', 'low', 'low', 'low', 'low', 'high', 'high', 'high', 'high', 'low', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'craniocerebral_injury+orthopedic_trauma_2_hours');

Проблема обчислювальної складності алгоритму індукції дерева рішень. Як вказано в роботі [Han, 2001], час виконання алгоритму індукції дерева рішень оцінюється величиною:

$$O(p \times \#(D) \times \log(\#(D))). \quad (4)$$

Тому нашою метою є підтвердити вищенаведений результат експериментально. Експеримент провели змінюючи кількість атрибутів p. Дерева рішень побудовані для кожного значення p.

Таблиця 1 – Залежність часу індукції дерева рішень від кількості атрибутів p (алгоритм на основі приросту інформації)

p	Час індукції дерева рішень, мс
1	122
2	252
3	392
4	450
5	550
6	762
7	820
8	900
9	1030
10	1080
11	1238
12	1282

Продолж. табл. 1

p	Час індукції дерева рішень, мс
13	1342
14	1424
15	1885
16	1402
17	1502
18	1585
19	2110
20	2212
21	3033

Таблиця 2 – Залежність часу індукції дерева рішень від кількості атрибутів p (алгоритм на основі відношення приростів інформації)

p	Час індукції дерева рішень, мс
1	100
2	290
3	610
4	740
5	832
6	930
7	1002
8	1242
9	1321
10	1472
11	1623
12	1724
13	1847
14	1981
15	2092
16	2595
17	1390
18	3033
19	2652
20	5096
21	5380

На рис.2 і 3 наведені оцінки часу індукції дерева рішень згідно (4).

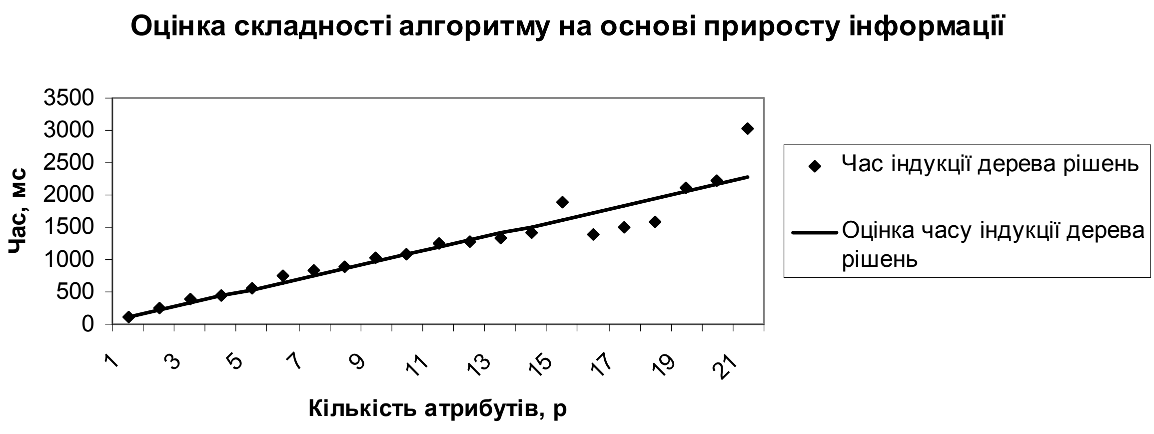


Рисунок 2

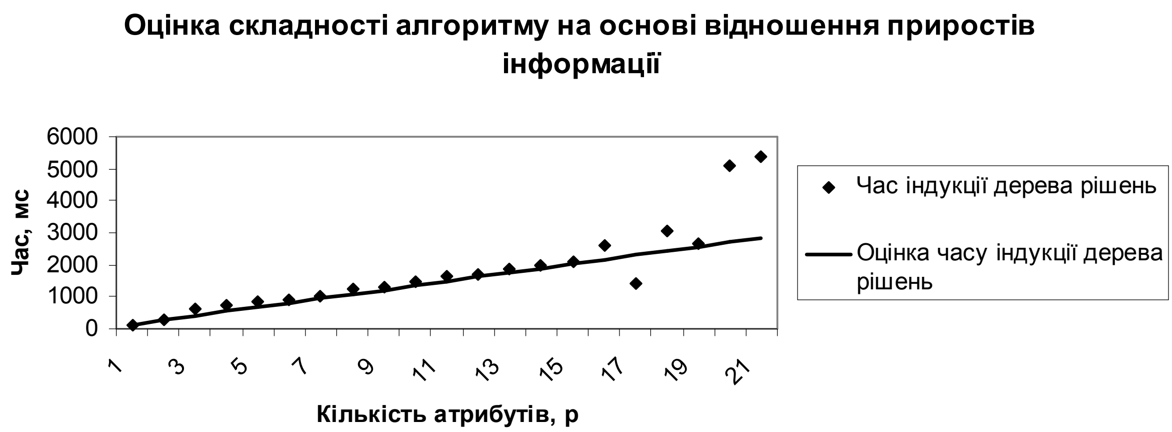


Рисунок 3

Список літератури

1. Соколов В.А. Множественные и сочетанные травмы / Соколов В.А. – ГЭОТАР, 2006г.
2. Han J. Data Mining: Concepts and Techniques / J. Han and M. Kamber – Morgan Kaufmann, San Francisco, 2001. – [1st ed.]
3. Hastie T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani and J.H. Friedman. – Springer, New York, 2001. – [1st ed.]
4. C.Ordonez, Comparing association rules and decision trees for disease prediction / C.Ordonez // Proc. ACM HIKM Workshop. – 2006. – P. 17-24.
5. Ordonez C. Integrating K-means clustering with a relational DBMS using SQL / C.Ordonez // IEEE Transactions on Knowledge and Data Engineering (TKDE). – 2006. – № 18(2). – P. 188-201.
6. Quinlan J.R. Induction of decision trees / J.R.Quinlan // Machine Learning. – 1986. – № 1. – P. 81-106.
7. Quinlan J.R. C4.5: Programs for Machine Learning / J.R.Quinlan. – Morgan Kaufmann, 1993.
8. Classification and Regression Trees / [L. Breiman, J. Friedman, R. Olshen, and C. Stone]. – Wadsworth International Group, 1984.
9. Марценюк В.П. О программной среде проектирования интеллектуальных баз данных / В.П. Марценюк, Н.О. Кравец // Клиническая информатика и телемедицина – 2004. – № 1. – С. 47-53.
10. Марценюк В.П. Математичні моделі в системі підтримки прийняття рішень страхового забезпечення лікування онкологічних захворювань: підхід на основі динаміки Гомперца / В.П. Марценюк, І.Є. Андрущак, І.С. Гвоздецька, Н.Я. Климук // Доповіді Національної академії наук України. – 2012. – № 10. – С. 34-39.
11. Марценюк В.П. Підхід на основі актуарних математичних моделей до задач страхової медицини / В. П. Марценюк, І.Є. Андрущак, Н.Я Климук // Медична інформатика та інженерія. Науково-практичний журнал. – 2010. – № 4. – С. 85-87.
12. Марценюк В.П. О модели онкологического заболевания со временем пребывания на стадии в соответствии с распределением Гомперца / В.П. Марценюк, Н.Я. Климук // Проблемы управления и информатики. Международный научно-технический журнал . – 2012. – № 6. – С. 137-143.
13. Марценюк В.П. Медична інформатика. Інструментальні та експертні системи / В.П. Марценюк , А.В. Семенець. – Тернопіль : Укрмедкнига, 2004. – 222 с.

References

1. Multiple and combined injuries – V.A. Sokolov, "GEOTAR" 2006.2. J..Han and M.Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 1st edition, 2001.
3. T.Hastie, R.Tibshirani and J.H.Friedman, The Elements of Statistical Learning, Springer, New York, 1st edition, 2001.
4. C.Ordonez, Comparing association rules and decision trees for disease prediction, In Proc. ACM HIKM Workshop, 2006, pp. 17-24.
5. C.Ordonez, Integrating K-means clustering with a relational DBMS using SQL, IEEE Transactions on Knowledge and Data Engineering (TKDE) 18(2) (2006), 188-201.

6. J.R.Quinlan. Induction of decision trees. Machine Learning, 1: 81-106, 1986.
7. J.R.Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
8. L.Breiman, J.Friedman, R.Olshen, and C.Stone. Classification and Regression Trees. Wadsworth International Group, 1984.
9. Martsenuk V.P., Kravets N.O. About the software environment of intelligent database / / Clinical informatics and telemedicine – 2004. – № 1. – P.47-53.
10. Martsenuk V.P. Matematichni modeli in sistemi pidtrimki of acceptance rishen insurance zabezpechennya likuvannya onkologichnih zahvoryuvan: pidhid on osnovi dinamiki Gompertz / V.P. Martsenyuk, I.Ye. Andrushchak, I.S. Gvozdetska, N.Y. Klymuk / / Dopovidi Natsionalnoї akademiї Sciences of Ukraine. – 2012. – № 10. – Pp. 34-39.
11. Martsenuk V.P. Pidhid on mathe osnovi actuarial models to problems strahovoї Medical / V.P. Martsenyuk, I.Ye. Andrushchak, N.Ya. Klymuk / / Medichna informatika that inzheneriya. Naukovyi impractical magazine. – 2010. – № 4. – S. 85-87.
12. Martsenuk V.P. On the model of cancer with a residence time on the stage, in accordance with the distribution of the Gompertz / V.P. Martsenyuk, N.Y. Klymuk / / Control and Informatics. International Science and Technology magazine. – 2012. – № 6. – S. 137-143.
13. Martsenuk V.P., A.V. Semenets Medichna informatika. Instrumentalni that ekspertni system. – Ternopil: Ukrmedkniga 2004. – 222.

RESUME

I.Ye. Andruschak, Y.S. Povstiana

Software Implementation By Induction of Decision Trees for Classification of Polytrauma: Computational Complexity

We consider the issue of development and program implementation of decision tree induction based on the information provided to construct a classification algorithm trauma.

Exploring in this example, the issue of computational complexity of decision tree induction algorithm found that:

– Time induction of decision tree based on the information provided is well approximated estimate (4) with a small number of attributes (in this case - to 15-16);

– With the number of attributes (in this case more than 15-16) while inducing decision trees, regardless of the choice of information measure attribute begins to deviate significantly from the estimates (4);

– With a small number of attributes induced by the decision tree constructed on the basis of growth rate information, or the ratio of growth are identical - that is, information measure, which is the basis of selection division attribute has no effect on the induced decision tree.

Prospects of this research is to analyze the performance of the software, depending on the amount of training data sets.

Стаття надійшла до редакції 05.04.2014.